DOCUMENT RESUME

ED 447 192		TM 032 080
AUTHOR	Kim, Jong-Pil	
TITLE	An Empirical Study of the Effect of Pooling Effect Sizes on Hedge's Homogeneity Test.	
PUB DATE	2000-04-00	
NOTE	18p.; Paper presented at the Annual Meeting of the American Educational Research Association (81st, New Orleans, LA, April 24-28, 2000).	
PUB TYPE	Reports - Evaluative (142) -- Speeches/Meeting Papers (150)	
EDRS PRICE	MF01/PC01 Plus Postage.	
DESCRIPTORS	Correlation; *Effect Size; *Meta Analysis; Simulation	
IDENTIFIERS	*Homogeneity Tests; *Type I Errors	

ABSTRACT

The homogeneity test provided by L. Hedges (1982) in meta-analysis has been widely used, mainly to test if the effect sizes share the same variance. Ignoring the intercorrelations among effect sizes affects the Type I error rate. The main purpose of this research was to study the impact of pooling effect sizes on the homogeneity test in effect size analyses. Simulations were conducted to study the effects of pooling effect sizes on Type I error rate and power of the Q test for varying sample size (N), number of studies (k), and proportion of pooling effect sizes (p) in the k studies. Simulation results show that composite meta-analysis seems to have smaller Type I error than typical meta-analysis. The difference in Type I error between typical and composite meta-analysis is relatively big when sample size or number of studies is large. Composite meta-analysis always has greater Type II error and smaller power than typical meta-analysis. These results mean that more caution is necessary when pooling effect sizes. (Contains 2 tables, 3 figures, and 16 references.) (SLD)

An Empirical Study of the Effect of Pooling Effect Sizes on Hedge's Homogeneity Test[1]

Jong-Pil Kim

Michigan State University

The meta-analytic techniques to synthesize related studies have been widely used in the social sciences. One of common approaches before estimating a mean effect size is to test if the effect sizes share a common population effect size. If the effect sizes do not share a common population effect size then sensitivity analyses are conducted to examine the influence of particular studies on combined effect size estimate. The homogeneity test provided by Hedges (1982) in meta-analysis has been widely used to test mainly if the effect sizes share same variance.

One typical feature of meta-analyses is treating multiple outcomes from single samples as if they were independent in calculating a grand mean effect size. Ignoring the intercorrelations among effect sizes affects the Type I error rate (Raudenbush, Becker, & Kalaian, 1988). The author's latest research showed that typical meta-analyses had a tendency of more significant results in homogeneity test, and categorical and regression analyses than when controlling dependent effect sizes (Kim, 1999). However, when correlation among dependent effect sizes is too low or when we are not sure if the effect sizes are dependent or not, then just combining or pooling effect sizes might bring some problems.

As mentioned above, in meta-analyses the effect-size analysis can involve two levels of statistical tests. Two types of decision-making errors in the first-stage homogeneity test can also affect the second-stage test of the magnitude of the common effect size. The main purpose of this research is to study the impact of pooling effect sizes on homogeneity test in effect size analyses.

[1] Paper presented at the annual meeting of the American Educational Research Association, New Orleans, April, 2000.

Nonindependence Issue

Landman and Dawes (1982) cautioned about five sources of nonindependence in meta-analysis. First, they cite multiple measures of outcomes from the same subjects within single studies; second, measures taken at multiple points in time from the same subjects (i.e., multiple occasions); third, nonindependence of scores within a single outcome measure; fourth, nonindependence of studies within a single article; and fifth, nonindependent samples across articles (p. 506-507). The second through fifth cases can be controlled by careful decision-making. For instance, when the same tests are examined several times in a study, only the last occasion could be selected (e.g., Kulik, Kulik, & Cohen, 1979). The third case happens when a study reports both a global index as well as more specific index, which is a part of the global index. In this case, choosing the specific index is ideal if it allows the study of interesting moderator variables. The fourth case occurs when some samples from two different experiments in a study are overlapping or the same. The same decision-making may be applied as used for the third case to arrive at independent samples. The last case may happen if the same sample appears in two different articles. In this case the more informative article should be selected. All four ways to treat nonindependence are used in this synthesis. However, the first case cannot be controlled by a decision-making, but by statistical consideration.

One common approach to the situation is for the meta-analyst to use all the statistics available in a particular study to calculate one mean effect size (Tracz, Elmore, & Pohlmann, 1992). The typical analysis then to treats each effect size from a given study as independent of the other effect sizes from the same study (e.g., Smith, Glass, & Miller, 1980). However, Glass, McGaw, and Smith (1981) recognized that "the data set to be analyzed [in a meta-analysis] will invariably contain complicated patterns of statistical dependence [since] each study is likely to yield more than one finding" (p.200). Bangert-Drowns (1986) stated, "multiple effect sizes from any one study cannot be regarded as independent and should not be used with statistical tests that assume their independence" (p. 397). In the same article (p. 392), he discussed the "Inflated Ns" problem. A report can have a greater influence on the meta-analytic findings if it uses many dependent measures. The "Inflated Ns" problem threatens the generalizability or external validity of a meta-analysis. Another problem is inflated Type I error (Raudenbush et al., 1988). Strube (1983) mentioned a general rule, that is, failure to adjust for nonindependence inflates the Type I error rate at the meta-analysis level.

Researchers have devised several methods for combining dependent data in meta-analysis. A strategy for reducing dependence of data is to select, on some predetermined basis,

a single dependent measure to represent each study (Cooper, 1979). But, the question "what is the best indicator among several dependent variables?" is too ambiguous. It is very difficult to make such a decision. A common strategy for dealing with studies that use multiple outcomes has been to average. This makes sense for providing a representative effect size estimate when the outcomes are parallel measures of a single construct (Raudenbush et. al., 1988). Instead of the mean, the median effect size is a more conservative option. Raudenbush et al. (1988) refer to this approach as study effect meta-analysis (p. 393) because of treating the study as the unit of analysis.

A statistical solution for this nonindependence problem within a study has been developed by Rosenthal and Rubin (1986). When the study has a big sample size and small differences of the intercorrelations between outcome measures, they suggest computing a composite effect size. Gleser and Olkin (1994) also showed how to calculate composite effect sizes within studies by using all individual intercorrelations among outcome variables. One difference between these two procedures is that Rosenthal and Rubin (1986) use a "typical" correlation, which is a correlation representative of all intercorrelations between the multiple measures.

One common feature of above approaches is calculating a representative effect sizes for dependent effect sizes. Combining dependent effect sizes to create one representative one for one measure from same sample seems to be reasonable. However, if the dependence is not certain, then just combining or pooling effect sizes may bring some problems for Type I error rate and power rate for statistical test in meta-analysis, mainly due to reduced sample sizes.

Q statistics

The biased effect size for each study is computed by

$$d_i = \frac{\left(\overline{X}_{iE} - \overline{X}_{iC}\right)}{S_i} \tag{1}$$

where $\overline{X}_{iC}$ and $\overline{X}_{iE}$ are the means in the ith study and S_i is the pooled standard deviation for study i calculated as:

$$S_i = \sqrt{\frac{(n_{iE}-1)(S_{iE})^2-(n_{iC}-1)(S_{iC})^2}{n_{iE}+n_{iC}-2}}, \tag{2}$$

where n_{iE} and n_{iC} are the number of each compared groups. The unbiased effect size (corrected for small sample bias) is calculated as

$$T_i = d_i \times \left[1 - \frac{3}{\{4 \times (n_i - 2) - 1\}} \right], \tag{3}$$

where $n_i = n_{iE} + n_{iC}$, with the conditional variance

$$v_i = \frac{ni}{n_{iE} \times n_{iC}} + \frac{d_i^2}{2n_i} \tag{4}$$

(Hedges & Olkin, 1985). The sample size varies across studies. Since estimates from the larger studies are more precise than estimates from smaller studies, larger studies are given more weight when the effects are pooled. The weight $w_i = 1/v_i$ is used. A pooled effect, or weighted mean effect ($T.$) can be calculated as:

$$T. = \frac{\sum_{i=1}^{k} (w_i T_i)}{\sum_{i=1}^{k} w_i}, \tag{5}$$

with a variance

$$v. = \frac{1}{\sum_{i=1}^{k} (\frac{1}{v_i})} \tag{6}$$

(Shadish & Haddock, 1994, p.265). In order to determine whether the studies can reasonably be described as sharing a common effect size, the following statistical test for homogeneity of effect size was performed (Hedges & Olkin, 1985):

$$Q = \sum_{i=1}^{k} \frac{(T_i - T.)^2}{v_i}. \tag{7}$$

The test statistic Q has an asymptotic chi-square distribution with $k-1$ degrees of freedom. When test statistic Q is greater than the critical value with k-1 degree of freedom, it is determined that the synthesis has heterogeneous data.

If the δ_i as a parameter effect sizes are not equal then Q has a noncentral chi-square distribution $k-1$ degree of freedom and noncentrality parameter (Chang, 1992).

$$\lambda = \sum_i \left(\frac{\delta_i - \delta.}{\sigma^2_{\delta_i}} \right)^2$$

Methods

Based on the main question "What are the effects of pooling effect sizes on Type I error rate and power of the Q test for varying study sample size (N), number of studies (k), and proportion of pooling effect sizes (p) in the k studies?" following procedures were implemented.

Simulation factors

The factors and their values reflect those of Harwell (1995), Chang (1992), and Hedges and Olkin (1985). Based on normal distribution of scores, numbers of effect sizes modeled were k = 5, 10, and 30 with group sample sizes of 10-10, 30-30 and an extreme value of 300-300. A reason of including one extreme sample size of 300 was to see its particular tendency about Type I error rate. In fact, there are many primary studies that contain more than 300 sample sizes in real setting. Unequal sample sizes were not included, but three different proportions of pooling effect sizes were considered. They were 20, 40 and 60% of the whole effect sizes. For instance, 2 effect sizes were assumed to correlate (i.e., came from same sample of one primary study) out of 5 studies when 40% proportion was used. Thus, the number of pooling effect sizes varied across different study sizes. These proportion of pooling effect sizes and studies were based on the author's last study (Kim, 1999). Pooling effect sizes were considered came from one, two, or three samples (primary studies) for each k = 5, 10 (except 20% proportion case), and 30 respectively. The noncentrality patterns were shown in Figure 1 including other factors. Only one value of δ was used because of its simplicity and middle size considering Chang (1992) and Harwell (1995).

Insert Figure 1 about here

Data generation

The data generation was done using MATLAB 5.3. The procedures were as follows. (A) Sample effect sizes were obtained from noncentral t statistics, computed using normal deviates and chi-squared random numbers generated. (B) Constants equal to the specified delta values were added to scores to create the desired noncentrality pattern as way of Chang (1992). The formula used for part (A) and (B) is as follow. (a) Got a normal deviate (Z_i). (b) Multiplied Zi by $\sqrt{(N_i / n_i \times n_i)}$. (c) Added noncentrality pattern of δ. (d) And divided it by $\sqrt{Ch_i / df}$, where Ch_i is a chi-square random number (Chang, 1992). (C) For composite meta-analysis, mean effect size(s) for each specified studies was (were) gained before computing Q statistic. (D) The Q statistic was computed for the k effect sizes using equation (7) for each typical and composite effect sizes, (E) Step (A) to (C) were repeated 2000 times [the same number of replications employed by Harwell (1996), Chang (1992), and Hedges and Olkin (1985)] for each combination of simulation factors. The proportions of significant Q tests across the 2000 replication represented empirical type I error rates and power values from typical and composite meta-analysis were compared to see the effect of pooling of effect sizes on the Q test. Overall 3 (different proportion of pooling effect sizes; 2 when k = 5.) X 3 (number of studies) X 3 (sample sizes) X 4 or 5 (sets of δ values) design was replicated.

Results

Adequacy of the simulation

One evidence of adequacy of the simulation is the mean effect sizes across the conditions studied. Table 1 shows mean effect sizes from 2000 simulation.

Insert Table 1 about here

For instance, all numbers in three first patterns of each proportion are pretty closer to zero, which indicates adequacy of the simulation. Another evidence is that Type I error rates and Power rates for typical meta-analysis are similar to theoretical rates. Those Type I error rates are close to 5.0 and delta pattern 2 and 5 (4 when k = 10) in 60% proportion possess pretty close values with theoretical values (See Table 2).

Insert Table 2 about here

Type I error rates of the Q test

The first rows of each proportion represents empirical error rates since delta pattern does not include any nonzero of δ. All of Type I error rates are pretty close to .05 for typical cases when sample sizes are big (30 and 300). Comparing Type I errors between typical and composite meta-analysis, composite meta-analysis always has smaller Type I error rates than typical case. This finding implies that pooling effect sizes is too conservative to reject the null hypothesis. One particular feature is that the difference of Type I error rates between two approaches is increasing when proportion of pooling effect sizes is increasing. This implies that the more pooling effect sizes a meta-analysis possesses, the more conservative the analysis is in rejecting the null hypothesis. Figure 2 presents this tendency clearly. Most cases show increasing Type I error rate difference (typical minus composite meta-analysis) across the proportion of pooling effect sizes.

Insert Figure 2 about here

Power of the Q test

When delta has not zeros, the sets estimate power values for the Q test. These values have similar patterns to the Type I error. Since most of power rate close to 1 when sample size is 300, it would be not explained from now. Power was largest for a given delta when n (fixing k) and k (fixing n) were larger. This tendency agrees with Chang (1992) and Harwell (1995), even if the values are a little different. Power for typical meta-analysis is always greater than composite meta-analysis. This means that composite meta-analysis underestimates statistical power that is supposed to. In other words, when independent effect sizes are considered as dependent, the statistical inference about Q statistic might overestimate Type II error and underestimate statistical power. Factors affecting the differences were detected with Figure 3.

Insert Figure 3 about here

When sample sizes, number of studies, and number of .5 of deltas in pooling effect sizes is bigger, the power difference (typical minus composite meta-analysis) between typical and composite meta-analysis is bigger. This indicates that under above conditions there are more possibilities for composite meta-analysis to overestimate Type II error and underestimate power of the Q test.

Conclusions

Composite meta-analysis seems to have smaller Type I error than typical meta-analysis. The difference in Type I error between typical and composite meta-analysis is relatively big when sample size and/or number of studies is big. This finding can be explained as follows. Composite meta-analysis is too conservative to reject the null hypothesis of homogeneity test. Then people tend to retain the null hypothesis when the alternative hypothesis is true. In turn, for example, people tend to use fixed effect model, to test if grand mean effect size is significant. Then, people tend to reject the null hypothesis of grand mean effect size test even if the alternative hypothesis is false because one uses smaller standard error than supposed to. Finally, people more likely commit Type I error for the test of grand mean effect sizes.

Harwell (1995) summarized that Chang (1992) suggested that meta-analysts should be concerned about the Type II errors since a Q test which was under-powered would lead to an unacceptably high probability of wrongly concluding that the model fits the data (p. 2). This study showed that composite meta-analysis always has greater Type II error and smaller power than typical meta-analysis. The difference in power rate between typical and composite meta-analysis is relatively big when sample size, number of studies, and number of .5 of deltas in pooling effect sizes is big. Thus more cautions are necessary when pooling (or combining) effect sizes.

In the future, a study to see the impact of dependence of effect sizes on Type I error & power rate of homogeneity test when ignoring the dependence might be pursued. For this, since the correlations among the dependent effect sizes might be taken account, Harwell (1995)'s previous data generation procedures and Gleser & Olkin (1994)'s formulas to generate dependent effect sizes are necessary to generate dependent effect sizes. In addition, it would be nicer if we can see the effect of dependence on random effect model of homogeneity too.

Reference

Bangert-Drowns, R.L. (1986). Review of developments in meta-analytic method. *Psychological Bulletin, 99(3), 388-399.*

Cooper, H. (1979). Statistically combining independent studies: A meta-analysis of sex differences in conformity research. *Journal of Personality and Social Psychology, 37*, 131-146.

Glass, G.V., McGaw, B., & Smith, M.L. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage.

Gleser, L.J., & Olkin, I. (1994). Stochastically dependent effect sizes. In H. Cooper & L. Hedges (Eds.) *The handbook of research synthesis (pp. 339-355).* New York: Russell Sage.

Chang, L. (1992). *A power analysis of the test of homogeneity in effect-sizes meta-analysis.* Unpublished doctoral dissertation. Michigan State University, East Lansing, MI.

Harwell, M. (1995). *An empirical study of the Hedges (1992) homogeneity test*. Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April).

Hedges, L.V. (1982). Fitting categorical models to effect sizes from a series of experiments. *Journal of Educational Statistics, 7, 119-137.*

Hedges, L.V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. New-York: Academic Press.

Kim, J.P. (1999). *Meta Analysis of Equivalence of Computerized and Pencil-and-Paper Testing for Ability Measures*. Paper presented at Annual Meeting of the Midwestern Educational Research Association (Chicago, IL, October)

Kulik, J.A., Kulik, C.C. & Cohen, P.A. (1979). A meta-analysis of outcome studies of Keller's personalized system of instruction. *American Psychologist, 34(4)*, 307-318.

Landman, J.T., & Dawes, R.M. (19982). Psychotherapy outcome: Smith and Glass' conclusions stand up under scrutiny. *American Psychologist, 37(5)*, 504-516.

Raudenbush, S.W., Becker, B.J., & Kalaian, H. (1988). Modeling multivariate effect sizes. *Psychological Bulletin, 103*, 111-120.

Rosenthal, R. & Rubin, D.B. (1986). Meta-analytic procedures for combining studies with multiple effect sizes. *Psychological Bulletin, 99(3)*, 400-406.

Smith, M.L., Glass, G.V., & Miller, T.I. (1980). *The benefits of psychotherapy*. Baltimore, MD: Johns Hopkins University Press.

Strube, M.J. (1985). Combining and comparing significance levels from nonsignificance hypothesis tests. *Psychological Bulletin, 97(2)*, 334-341.

Traze, S. M., Elmore, P.B. & Pohlmann, J.T. (1992) Correlational meta-analysis: independent and nonindependent. *Educational and Psychological Measurement, 52.* 879-888.

Table 1. Means of simulated effect sizes

	k*	5						10						30					
	n**	10		30		300		10		30		300		10		30		300	
pd***	delta****	t*****	c******	t	c	t	c	t	c	t	c	t	c	t	c	t	c	t	c
20%	1							0.007	-0.003	0.003	0.003	-0.001	0.001	0.000	0.002	0.000	0.000	0.000	0.000
	2							0.046	0.057	0.050	0.057	0.048	0.054	0.020	0.025	0.016	0.018	0.016	0.108
	3							0.090	0.075	0.095	0.080	0.098	0.082	0.046	0.035	0.046	0.034	0.049	0.036
	4							0.143	0.107	0.147	0.110	0.146	0.108	0.077	0.052	0.080	0.054	0.081	0.054
	5													0.111	0.072	0.112	0.071	0.115	0.073
40%	1	0.003	0.012	0.000	0.000	0.001	0.000	0.007	0.008	0.003	0.003	-0.001	-0.001	0.000	0.002	0.000	0.001	0.000	0.000
	2	0.098	0.120	0.097	0.121	0.122	0.098	0.046	0.057	0.050	0.062	0.048	0.060	0.020	0.021	0.016	0.026	0.016	0.023
	3	0.195	0.181	0.190	0.184	0.196	0.183	0.143	0.122	0.147	0.123	0.146	0.123	0.077	0.043	0.080	0.046	0.081	0.046
	4	0.280	0.236	0.293	0.244	0.296	0.256	0.235	0.178	0.240	0.179	0.026	0.184	0.142	0.071	0.144	0.069	0.147	0.070
	5													0.201	0.098	0.210	0.093	0.213	0.093
60%	1	0.003	0.005	0.000	0.000	0.001	0.001	0.007	-0.005	0.003	0.000	-0.001	0.001	0.000	0.000	0.000	0.000	0.000	-0.002
	2	0.098	0.078	0.097	0.082	0.122	0.081	0.046	0.078	0.050	0.082	0.048	0.081	0.020	0.033	0.016	0.032	0.016	0.033
	3	0.195	0.210	0.190	0.218	0.196	0.218	0.184	0.154	0.192	0.160	0.197	0.164	0.111	0.064	0.112	0.064	0.115	0.065
	4	0.334	0.242	0.345	0.248	0.347	0.247	0.334	0.242	0.345	0.248	0.347	0.247	0.208	0.103	0.210	0.098	0.212	0.098
	5	0.383	0.329	0.394	0.330	0.397	0.329							0.298	0.127	0.307	0.131	0.313	0.130

* k: Numer of studies
** n: Sample sizes
*** pd: Proportion of pooling effect sizes in a meta-analysis
**** delta: Patterns of deltas
***** t: Typical meta-analysis
****** c: Composite meta-analysis

Table 2. Type I error rate and power rate for the Q test

	k	5						10						30					
	n	10		30		300		10		30		300		10		30		300	
pd	delta	t	c	t	c	t	c	t	c	t	c	t	c	t	c	t	c	t	c
	1							4.2	2.9	4.6	2.5	5.3	4.2	3.1	1.7	4.9	3.5	4.5	2.7
	2							7.3	5.6	17.4	16.0	99.6	99.5	4.4	3.1	11.0	7.8	95.1	94.8
20%	3							9.5	6.3	30.8	18.0	100.0	99.9	7.9	4.8	28.6	15.2	100.0	100.0
	4							12.8	8.5	42.1	29.2	100.0	100.0	11.6	5.6	52.1	24.3	100.0	100.0
	5													15.0	7.5	64.3	36.5	100.0	100.0
	1	4.4	2.7	4.8	2.7	5.3	3.2	4.2	2.4	4.6	2.8	5.3	3.1	3.1	1.9	4.9	1.7	4.5	2.2
	2	8.8	7.4	22.9	22.7	99.6	99.9	7.3	4.5	17.4	14.8	99.6	99.7	4.4	2.5	11.0	6.6	95.1	96.1
40%	3	10.3	7.9	34.5	19.8	100.0	99.6	12.8	6.0	42.1	25.0	100.0	100.0	11.6	4.2	51.1	15.4	100.0	100.0
	4	12.0	9.7	34.4	29.0	100.0	100.0	13.4	8.2	51.2	37.2	100.0	100.0	16.6	4.8	73.0	23.9	100.0	100.0
	5													22.3	5.6	83.3	37.2	100.0	100.0
	1	4.4	2.4	4.8	2.5	5.3	2.2	4.2	2.0	4.6	2.3	5.3	2.2	3.1	1.7	4.9	1.4	4.5	1.5
	2	8.8	7.3	22.9	21.4	99.6	99.9	7.3	4.5	17.4	16.2	99.6	99.6	4.4	2.5	11.0	8.4	95.1	98.2
	(T)	9.9		23.3		99.6		8.4		18.3		99.3		6.8		12.0		95.0	
60%	3	10.3	9.9	34.5	17.0	100.0	98.1	13.4	6.2	49.5	25.9	100.0	100.0	15.0	2.9	64.3	13.7	100.0	100.0
	4	12.0	6.5	34.4	17.6	100.0	98.1	12.5	7.4	42.5	30.7	100.0	100.0	20.7	4.5	83.7	25.3	100.0	100.0
	5	8.9	7.6	23.5	21.5	99.7	99.8							20.5	6.3	81.8	35.9	100.0	100.0
Theoretical values		10.0		23.7		99.7		14.3		43.2		100.0		25.7		82.0		100.0	

	pd* = 20%					pd = 40 %					pd = 60 %				
delta**	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
k***=5						k composite**** = 4					k composite = 3				
1						0	0	0.5	0.5		0	0	0.5	0.5	0.5
2						0	0	0	0.5		0	0	0	0.5	0.5
3						0	0	0	0		0	0	0	0	0.5
4						0	0	0	0		0	0	0	0	0
5						0	0.5	0.5	0.5		0	0.5	0.5	0.5	0.5
k=10	k composite = 9					k composite = 8					k composite = 6				
1	0	0	0.5	0.5		0	0	0.5	0.5		0	0	0.5	0.5	
2	0	0	0	0.5		0	0	0.5	0.5		0	0	0.5	0.5	
3	0	0	0	0		0	0	0	0.5		0	0	0.5	0.5	
4	0	0	0	0		0	0	0	0.5		0	0	0	0.5	
5	0	0	0	0		0	0	0	0		0	0	0	0.5	
6	0	0	0	0		0	0	0	0		0	0	0	0.5	
7	0	0	0	0		0	0	0	0		0	0	0	0	
8	0	0	0	0		0	0	0	0		0	0	0	0	
9	0	0	0	0		0	0	0	0		0	0	0	0	
10	0	0.5	0.5	0.5		0	0.5	0.5	0.5		0	0.5	0.5	0.5	
k=30	k composite = 27					k composite = 21					k composite = 15				
1	0	0	0.5	0.5	0.5	0	0	0.5	0.5	0.5	0	0	0.5	0.5	0.5
2	0	0	0.5	0.5	0.5	0	0	0.5	0.5	0.5	0	0	0.5	0.5	0.5
3	0	0	0	0.5	0.5	0	0	0.5	0.5	0.5	0	0	0.5	0.5	0.5
4	0	0	0	0.5	0.5	0	0	0.5	0.5	0.5	0	0	0.5	0.5	0.5
5	0	0	0	0	0.5	0	0	0	0.5	0.5	0	0	0.5	0.5	0.5
6	0	0	0	0	0.5	0	0	0	0.5	0.5	0	0	0.5	0.5	0.5
7	0	0	0	0	0	0	0	0	0.5	0.5	0	0	0	0.5	0.5
8	0	0	0	0	0	0	0	0	0.5	0.5	0	0	0	0.5	0.5
9	0	0	0	0	0	0	0	0	0	0.5	0	0	0	0.5	0.5
10	0	0	0	0	0	0	0	0	0	0.5	0	0	0	0.5	0.5
11	0	0	0	0	0	0	0	0	0	0.5	0	0	0	0.5	0.5
12	0	0	0	0	0	0	0	0	0	0.5	0	0	0	0.5	0.5
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.5
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.5
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.5
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.5
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.5
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.5
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
30	0	0.5	0.5	0.5	0.5	0	0.5	0.5	0.5	0.5	0	0.5	0.5	0.5	0.5

Figure 1. Pattern of pooling effect sizes with values of deltas

* pd: proportion of pooling effect sizes in a meta-ana;ysis
** delta: Patterns of deltas
*** k: Numer of studies
**** k composite: Number of studies after pooling effect sizes

Figure 2. Type I error difference between typical and composite meta-analysis

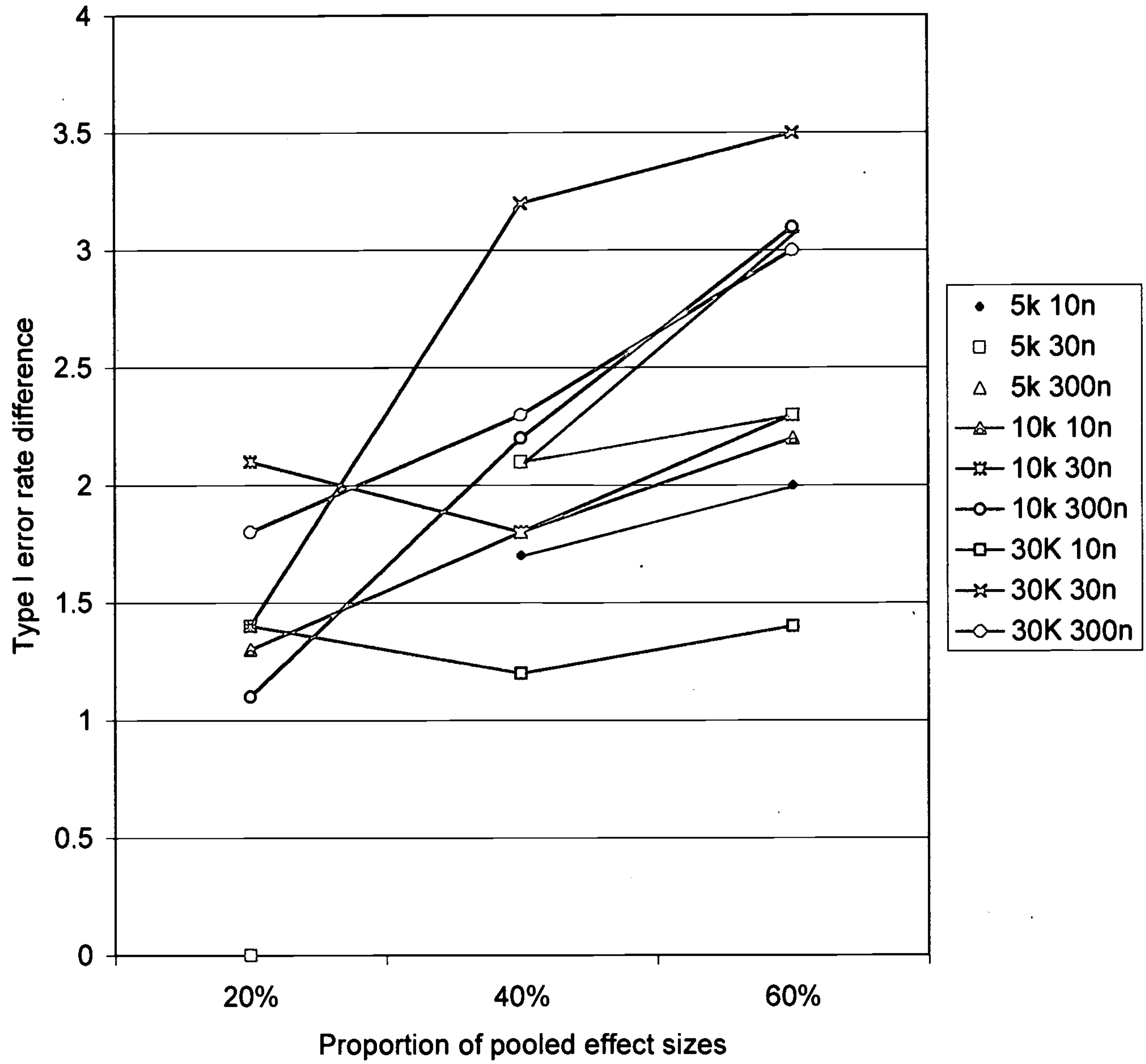
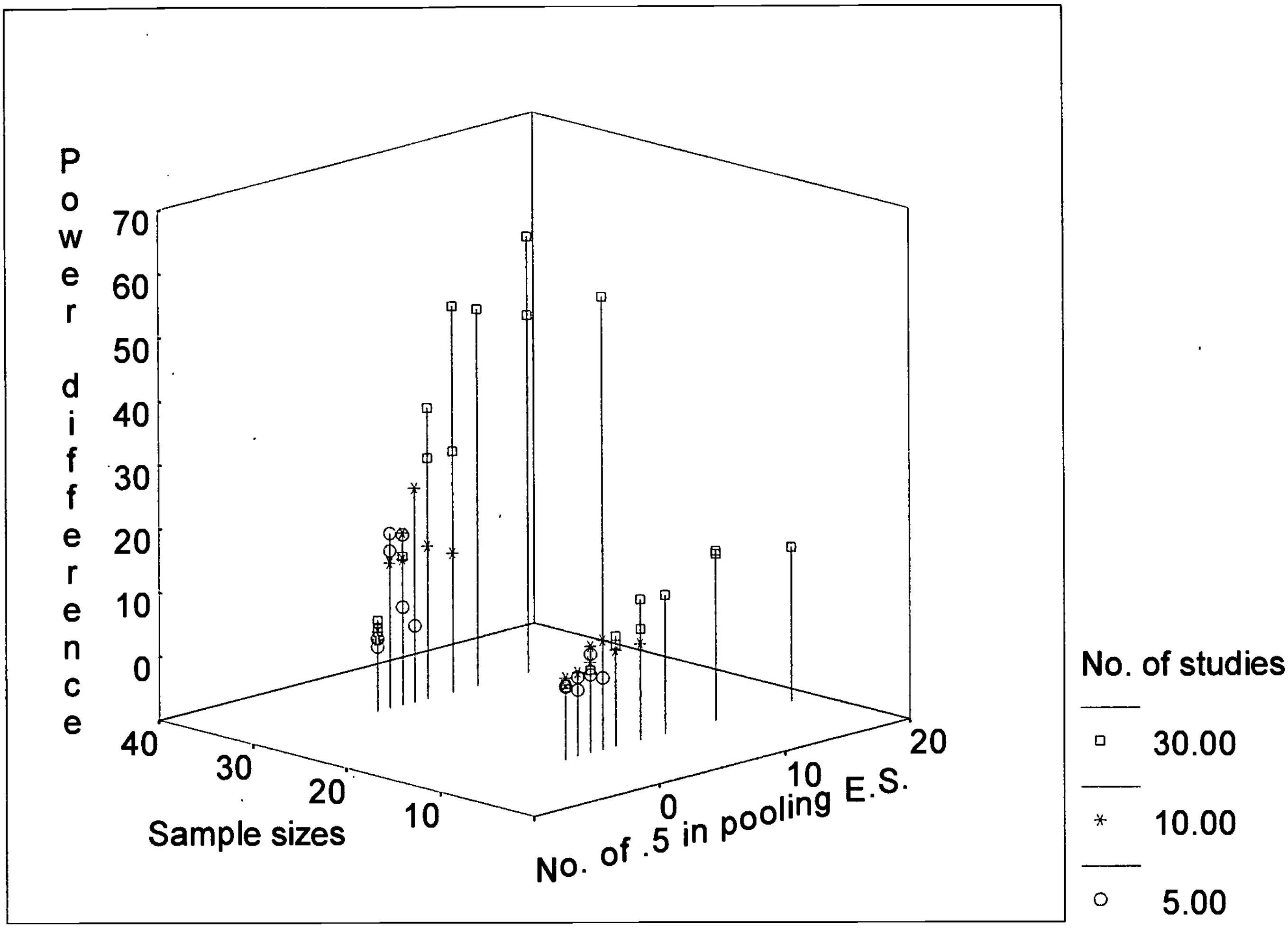


Figure 3. Power difference between typical and composite meta-analysis



Abstract

One typical feature of meta-analyses is treating multiple outcomes from single samples as if they were independent in calculating a grand mean effect size. Ignoring the intercorrelations among effect sizes affects the Type I error rate. However, when correlation among dependent effect sizes is too low or when we are not sure if the effect sizes are dependent or not, then just combining or pooling effect sizes might bring some problems. The main purpose of this research is to study the impact of pooling effect sizes on homogeneity test in effect size analyses. Based on the main question "What are the effects of pooling effect sizes on Type I error rate and power of the Q test with 3 sample sizes, 3 number of studies, 3 proportion of pooling effect sizes in the k studies, and 4 or 5 kinds of noncentrality patterns?", 2000 replications were implemented.

Composite meta-analysis seems to have smaller Type I error than typical meta-analysis. The difference in Type I error between typical and composite meta-analysis is relatively big when sample size and/or number of studies is big. This finding implies that composite meta-analysis is too conservative to reject the null hypothesis of homogeneity test, in turn, has more likely higher Type I error and lower power for the test of grand mean effect sizes. This study also showed that composite meta-analysis always has greater Type II error and smaller power than typical meta-analysis. The difference in power rate between typical and composite meta-analysis is relatively big when sample size, number of studies, and number of .5 of deltas in pooling effect sizes is big. This results recommend that more cautions are necessary when pooling (or combining) effect sizes.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC®

TM032080

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: An Empirical Study of the effect of pooling effect sizes on Hedge's homogeneity test.	
Author(s): Jong-Pil Kim	
Corporate Source: 2000. AERA. New Orleans.	Publication Date:

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 1	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2A	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2B
Level 1 ☑	Level 2A ☐	Level 2B ☐
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, → please

Signature: [signature]	Printed Name/Position/Title: Kim, Jong Pil / Dotoral candidate.	
Organization/Address: Michigan State U. E. Lansing MI 48823	Telephone: 517-882-6375	FAX:
	E-Mail Address: Kimjongp@pilot.msu.edu.	Date: 11. 1. 00

(over)

ERIC® Clearinghouse on Assessment and Evaluation

University of Maryland
1129 Shriver Laboratory
College Park, MD 20742-5701

Tel: (800) 464-3742
(301) 405-7449
FAX: (301) 405-8134
ericae@ericae.net
http://ericae.net

May 8, 2000

Dear AERA Presenter,

Hopefully, the convention was a productive and rewarding event. As stated in the AERA program, presenters have a responsibility to make their papers readily available. If you haven't done so already, please submit copies of your papers for consideration for inclusion in the ERIC database. We are interested in papers from this year's AERA conference and last year's conference. If you have submitted your paper, you can track its progress at http://ericae.net.

Abstracts of papers accepted by ERIC appear in *Resources in Education (RIE)* and are announced to over 5,000 organizations. The inclusion of your work makes it readily available to other researchers, provides a permanent archive, and enhances the quality of *RIE*. Abstracts of your contribution will be accessible through the printed and electronic versions of *RIE*. The paper will be available through the microfiche collections that are housed at libraries around the world and through the ERIC Document Reproduction Service.

We are gathering all the papers from the **2000 and 1999 AERA Conference**. We will route your paper to the appropriate clearinghouse. You will be notified if your paper meets ERIC's criteria for inclusion in *RIE*: contribution to education, timeliness, relevance, methodology, effectiveness of presentation, and reproduction quality.

Please sign the Reproduction Release Form enclosed with this letter and send **two** copies of your paper. The Release Form gives ERIC permission to make and distribute copies of your paper. It does not preclude you from publishing your work. You can mail your paper to our attention at the address below. Please feel free to copy the form for future or additional submissions.

Mail to: AERA 2000/ERIC Acquisitions
University of Maryland
1129 Shriver Laboratory
College Park, MD 20742

Sincerely,

Lawrence M. Rudner

Lawrence M. Rudner, Ph.D.
Director, ERIC/AE